Week 3 Report: Hypothesis Testing

Data Analytics & Machine Learning Research Lab

Vo Nguyen Le Duy

Tran Minh Thien

23rd January 2025

Contents

1 Hypothesis testing $\bar{u}$

1.1 Problem

- Given two sets of values, X and Y, each containing 50 elements, perform a hypothesis testing to determine whether their means are equal.
- The **Null hypothesis** (H_0) is:

$$\mu_x = \mu_y$$

- The **Alternative Hypothesis** (H_1) is:

$$\mu_x \neq \mu_y$$

1.2 Analysis

- Because we compare 2 sets of values with 50 elements each, we will use Z-Test. Specifically, we compare the mean value of 2 separate sets so we use the 2-Sample Z-Test:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

  $\bar{x}, \bar{y}$: Mean value of the two sets

  σ_x^2, σ_y^2 : Variance of the two sets

  n_x, n_y : Size of the two sets

- Calculate variance:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Distribution of Z value: **Gaussian Distribution**

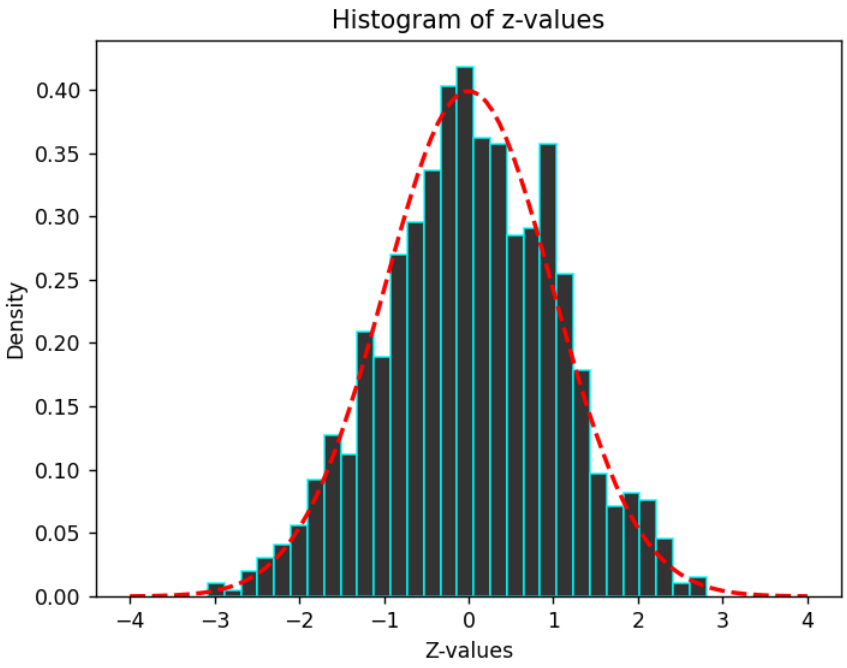

Figure 1: Z Distribution

- **P_value graph:**

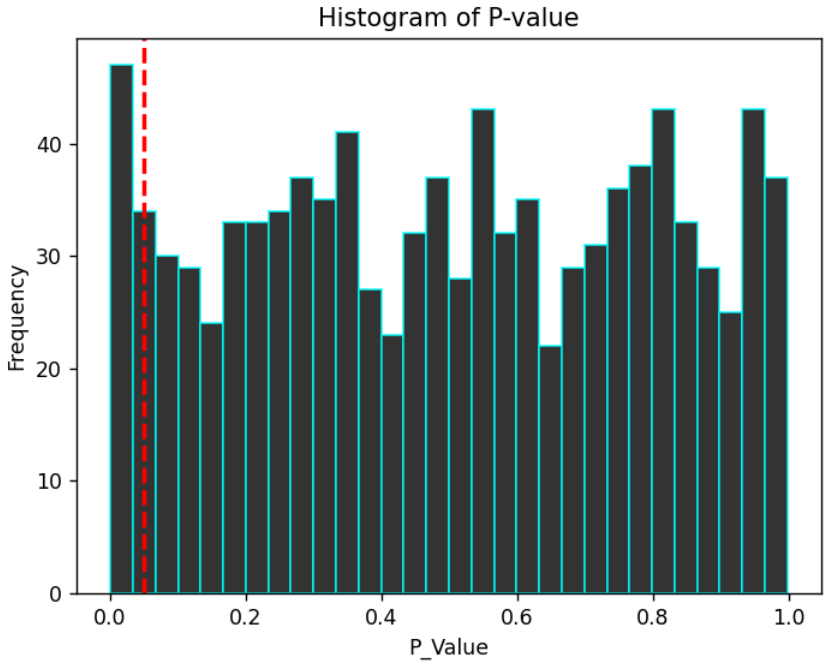

Figure 2: P_value

1.3 Coding

2 Hypothesis testing Linear Regression

2.1 Problem

- Given a linear regression model, use hypothesis testing to determine whether the coefficients in a model are different from 0 (**Important**) or equal 0 (**Unnecessary**)
- The **Null hypothesis** (H_0) is:

$$\beta_i = 0$$

- The **Alternative hypothesis** (H_1) is:

$$\beta_i \neq 0$$

2.2 Analysis

- We use Z-Test to calculate p-value

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

- Calculate **Residuals**: $\hat{\beta}_i = \hat{\omega}_i - 0$
- Calculate **Standard Error**: $SE(\hat{\beta}_i) = \sqrt{Cov(\hat{\beta}_i)}$

  Covariance Matrix:

$$Cov(\hat{\beta}) = \sigma^2 \left(X^T X\right)^\dagger$$

  Variance:

$$\frac{RSS}{n-k}$$

  n: number of elements

  k: number of features + bias

RSS (Residuals Sum of Square):

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Calculate **P-value**:

$$P\_value = 2 \ * \ \min\left[1 - \Phi(|z_i|),\ \Phi(|z_i|)\right]$$

$\Phi(x)$: CDF

- **Coefficients**: [1, 2, 0, 0.8, 0, 1.5]
- **Distribution: Gaussian Distribution**

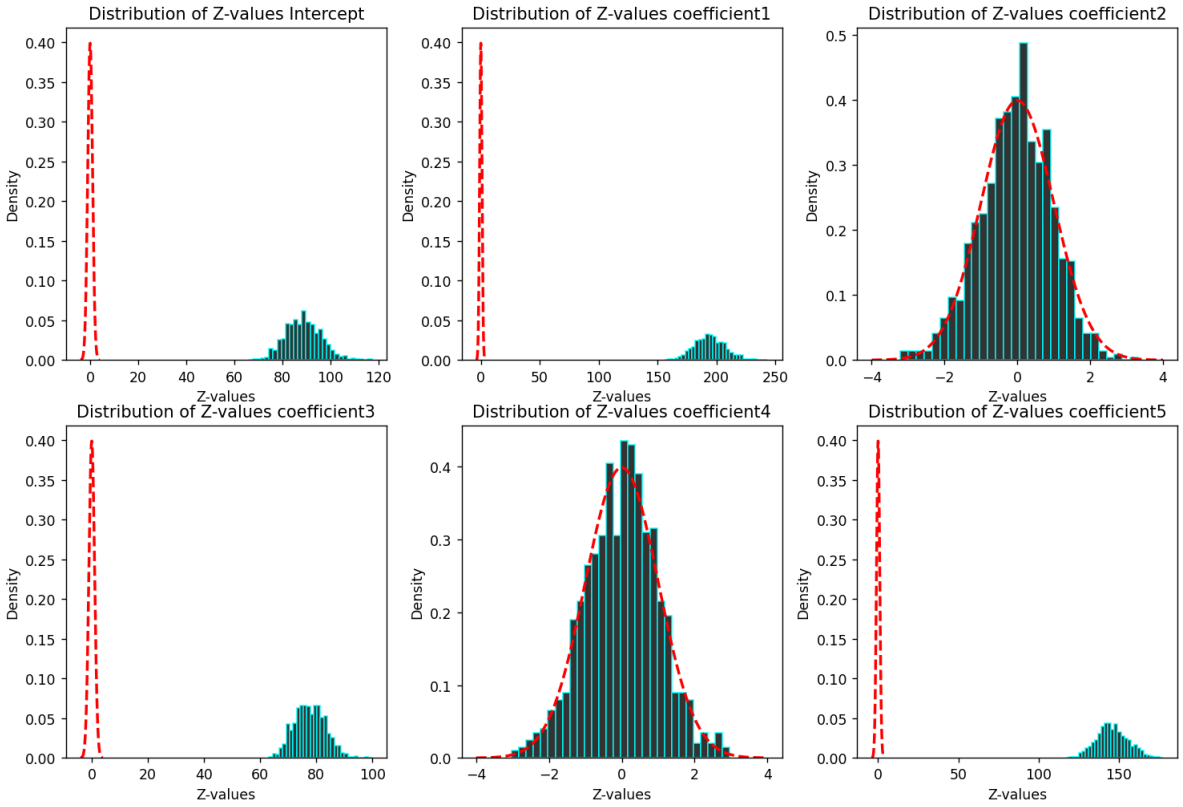

Figure 3: Linear Regression Z-Value

- **P_value Graph**

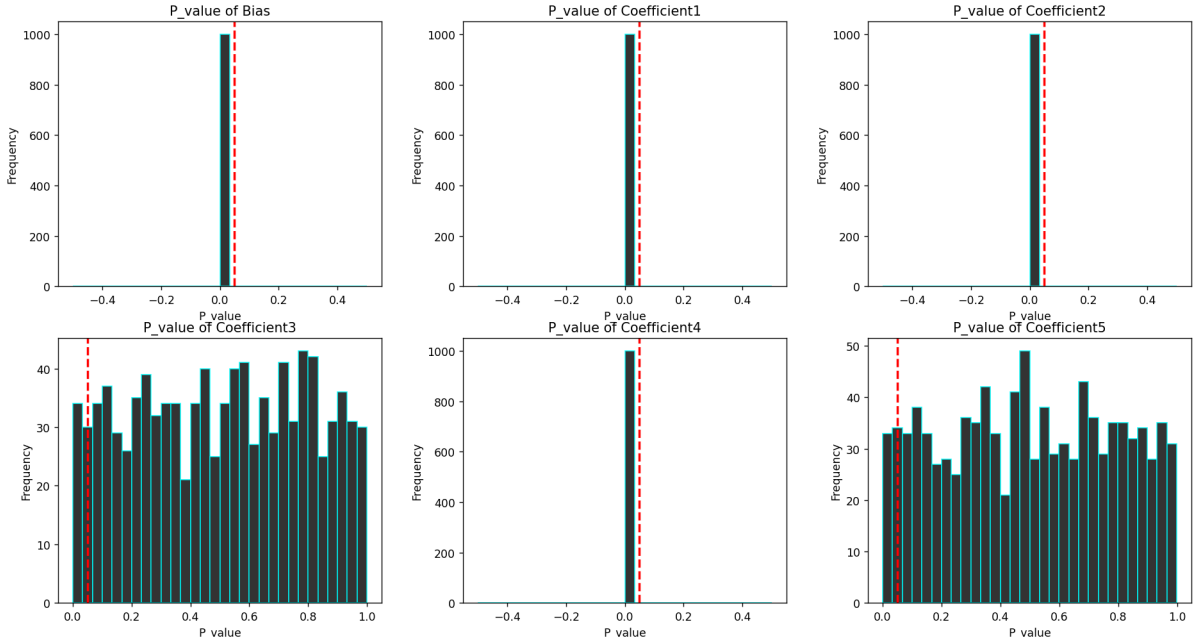

Figure 4: Linear Regression P_value

2.3 Code